Published online 5 November 2013 *Nucleic Acids Research, 2014, Vol. 42, Database issue* ***D625–D632***
doi:10.1093/nar/gkt1094

POGO-DB—a database of pairwise-comparisons of genomes and conserved orthologous genes

Yemin Lan[1], J. Calvin Morrison[2], Ruth Hershberg[3,*] and Gail L. Rosen[2,*]

[1]School of Biomedical Engineering, Science and Health Systems, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA, [2]Electrical & Computer Engineering Department, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA and [3]Rachel & Menachem Mendelovitch Evolutionary Processes of Mutation & Natural Selection Research Laboratory, Department of Genetics, the Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa 31096, Israel

Received August 15, 2013; Revised October 15, 2013; Accepted October 16, 2013

ABSTRACT

POGO-DB (http://pogo.ece.drexel.edu/) provides an easy platform for comparative microbial genomics. POGO-DB allows users to compare genomes using pre-computed metrics that were derived from extensive computationally intensive BLAST comparisons of >2000 microbes. These metrics include (i) average protein sequence identity across all orthologs shared by two genomes, (ii) genomic fluidity (a measure of gene content dissimilarity), (iii) number of 'orthologs' shared between two genomes, (iv) pairwise identity of the 16S ribosomal RNA genes and (v) pairwise identity of an additional 73 marker genes present in >90% prokaryotes. Users can visualize these metrics against each other in a 2D plot for exploratory analysis of genome similarity and of how different aspects of genome similarity relate to each other. The results of these comparisons are fully downloadable. In addition, users can download raw BLAST results for all or user-selected comparisons. Therefore, we provide users with full flexibility to carry out their own downstream analyses, by creating easy access to data that would normally require heavy computational resources to generate. POGO-DB should prove highly useful for researchers interested in comparative microbiology and benefit the microbiome/metagenomic communities by providing the information needed to select suitable phylogenetic marker genes within particular lineages.

INTRODUCTION

The ever-increasing number of available fully sequenced microbial genomes opens up new avenues of research into microbial evolution. Yet these new opportunities also come with new challenges. Analyses and comparison of whole-genome sequences most often require the performance of whole-genome comparisons using tools such as BLAST (1–4). The computational cost of such comparisons increases as a function of the number of genomes that need to be compared. For example, if one is interested in studying the *Escherichia*/*Shigella* lineage that is currently represented by 54 fully sequenced genomes in the National Center for Biotechnology Information (NCBI) whole-genome database (as of July 2012), one would have to conduct 1431 two-way whole-genome BLAST comparisons, which would take several hundred CPU-hours. With >2000 genomes currently available, the all-against-all BLAST comparisons become extremely resource-consuming (>150 000 CPU-hours using the advanced digital resources and services provided by XSEDE). Here, we provide a resource that compiles pre-computed BLAST comparisons of all protein-coding genes against all protein-coding genes for pairs of genomes. The resource allows the user to compare genomes using a set of different metrics (that we calculated based on these BLAST comparisons), including overall protein sequence identity, genome fluidity (a measure of dissimilarity in gene content) and number of shared orthologs. The resource also allows for the download of raw BLAST results, saving users the need to re-perform these comparisons using their own resources. Users can select any group or groups of genomes they want to compare and conduct both inter-group and intra-group comparisons. POGO should therefore prove useful for anyone interested in comparing microbial genomes and/or studying genome evolution.

In addition to its general usefulness for comparative biology of microbes, the POGO database has also been designed with a special application in mind, namely, the identification of phylogenetic marker genes for specific

*To whom correspondence should be addressed. Tel: +972 4 8295282; Fax: +972 4 8295225; Email: ruthersh@tx.technion.ac.il
Correspondence may also be addressed to Gail Rosen. Tel: +1 215 895 0400; Fax: +1 215 895 1695; Email: gailr@ece.drexel.edu

*Equal contributions as last author

lineages that may be of interest in various metagenomic studies. Metagenomic approaches allow for the characterization of whole communities of microbes within natural and host environments. A key aim of metagenomic studies is to infer the phylogenetic composition of a considered microbial community (5). Often in metagenomic studies, a phylogenetic marker gene is amplified, sequenced and used to infer phylogeny. The by far most commonly used marker gene is the 16S ribosomal RNA (rRNA) gene (6–9). Sequencing of the nearly 1600-bp 16S rRNA gene can produce a community census with less sequencing than whole-genome sequencing of million-bp genomes (or it can allow deeper sequencing for the same cost). This makes 16S rRNA amplicon sequencing attractive for the purpose of characterizing the population of a metagenome. The 16S rRNA gene has proven to be particularly useful in inferring phylogenetic composition because it is universally present across microbes, contains both universally conserved and variable regions and is thought not to undergo much horizontal gene transfer (10). Owing to its great usefulness, comprehensive databases cataloging 16S rRNA sequences across the prokaryotic spectrum have been built (11–13). Yet, use of the 16S rRNA gene carries certain limitations. Importantly, the 16S RNA gene is known to be a poor marker for the inference of phylogeny within closely related lineages (14,15). Thus, if one wishes to conduct a higher resolution phylogenetic analysis of their community, 16S rRNA may not be the best marker to use.

Several recent studies have started looking into using markers other than the 16S rRNA gene to infer phylogeny within metagenomes, such as *rpoB, amoA, pmoA, nirS, nirK, nosZ* and *pufM* (16–18). Housekeeping genes have been suggested to be useful for discriminating lineages, as they are a major component of the core genes for a lineage (19) and are thought to have less environmental pressure than other genes (20). Wu *et al*. identified 31 housekeeping genes from 100 genomes to be used to reconstruct phylogeny (21), which has also been used to speed up the taxonomic classification (22). POGO-DB allows users to easily examine the relative evolutionary rates of the different markers and to investigate how well the percentage identities of different markers correlate to overall amino acid identities across the genomes of interest (which should provide the most reliable estimates of phylogenetic relatedness). These data should be useful for choosing marker genes for metagenomic studies.

GENOME SIMILARITY METRICS PROVIDED BY POGO-DB

POGO-DB comprises four types of metrics useful for the study of different aspects of genome similarity: (i) the average amino acid identity of all orthologous proteins within a pair of genomes (AAI) (23). Pairs of genomes that are more closely related are expected to have higher AAI than more distantly related genome pairs. AAI combines information across all shared proteins within a pair of genomes and is therefore expected to provide much more reliable estimation of the degree of genetic relatedness between a pair of genomes, compared with estimates gathered using degree of similarity for a single locus of a small number of loci; (ii) genomic fluidity, a measure of gene content dissimilarity that estimates the proportion of genes unique to each of the two genomes considered, of the total number of genes contained within these two genomes (24); (iii) the 'number of orthologs' as defined by two criteria (see later); and (iv) percentage identities of 74 potential marker genes that we found to be present in at least 90% of prokaryotes (including the 16S rRNA gene).

The acquisition of POGO data is illustrated in Figure 1. We downloaded all complete prokaryote genomes available from NCBI (as of July 2012) and extracted their 16S rRNA genes. The 16S rRNA genes were aligned using the Needleman–Wunsch algorithm (25). For each pair of genomes whose 16S rRNA gene identity was >80%, we conducted a genome-wide two-way BLAST (26) of all protein sequences against all protein sequences [GenBank (27) annotated CDS were used]. The resulting reciprocal best hits were aligned using the Smith–Waterman algorithm (28) to determine their levels of sequence identity more accurately.

When estimating the average protein identity across orthologs, we wanted to be strict in defining orthologs, so as not to use false ortholog pairs that could skew our results. Therefore, to qualify as an orthologous pair, we required that the reciprocal best hits align across at least 70% of the shorter sequence, with at least 30% amino acid identity [as in (23)]. The average AAI of a genome pair was then calculated as the average amino acid identity across all identified ortholog pairs.

To compute genomic fluidity (a measure of gene content dissimilarity), we wanted to be more conservative with regard to inferring gene absence. Therefore, we used a less strict threshold for inferring whether a gene from one genome was present in the other. To say that a gene from one genome was present in the other, the BLAST best hit would have to align across at least 50% of the shorter sequence with at least 10% amino acid identity. Genomic fluidity is then computed as:

$$\frac{U1+U2}{M1+M2},$$

where *U1* and *U2* are the number of genes unique to each of the two genomes, and *M1* and *M2* are the total number of genes contained in each of the two genomes [as in (24)].

POGO-DB provides the number of orthologs identified for each genome pair using both the more and less strict thresholds (alignment across at least 70% of sequence with at least 30% identity, versus alignment across at least 50% of sequence with at least 10% identity). For the computation of both the average AAI and genomic fluidity, a pair of genomes should have at least 200 orthologs to be included.

POGO-DB also provides the pairwise percentage identity calculated for 74 marker genes (including the 16S rRNA gene). These marker genes are targeted because they are universal among genomes in the COG database (29), and we later identified them as being

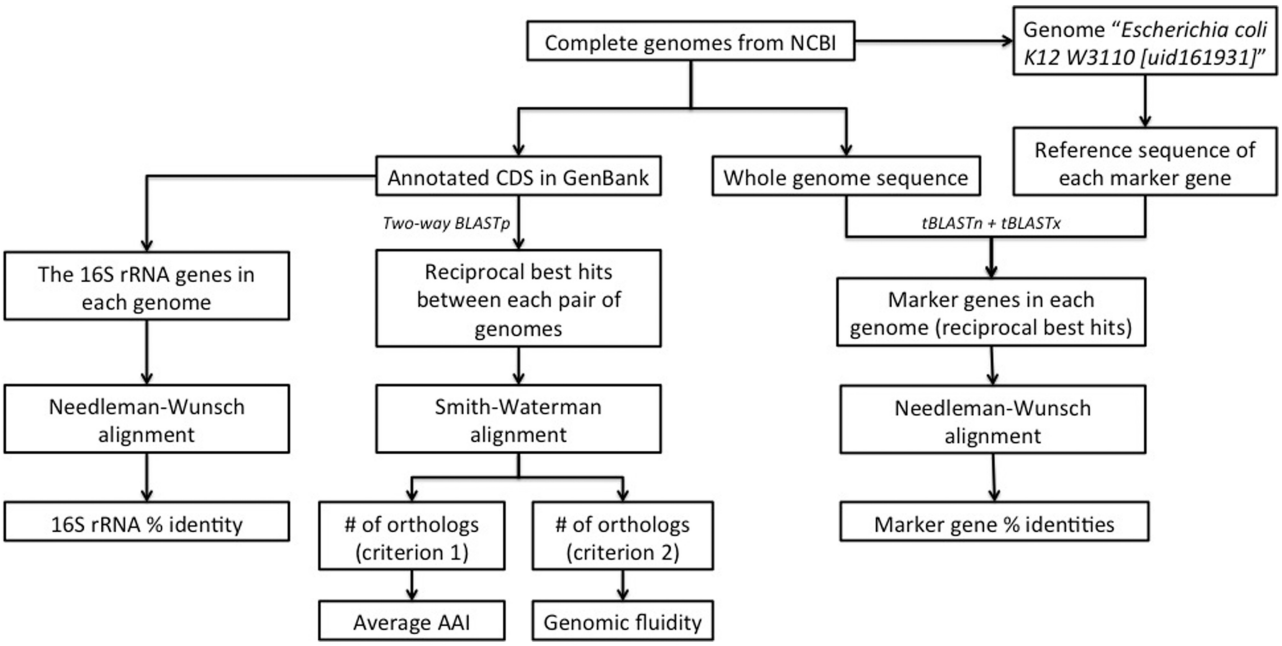

Figure 1. Data acquisition flowchart shows the general process of POGO data generation.

conserved in at least 90% of prokaryotes. The 16S rRNA genes in each genome were directly extracted from the GenBank annotation. Because annotations are not always consistent and gene names can vary across genomes, the sequences of the other marker genes were acquired by BLAST-ing a set of reference marker genes (annotated in genome *Escherichia coli K12 W3110 [uid161931]*) against each genome. For each marker gene (including the 16S rRNA gene), Needleman–Wunsch alignment is applied to acquire the percentage identity between each pair of genomes. Here, the percentage identities were calculated at the DNA rather than the protein level (to be consistent with the 16S rRNA comparisons).

LAYOUT AND USE OF POGO-DB

Figure 2 shows the layout of POGO database. POGO allows users to assign any combination of genomes, species and genera into two groups. While the search of genome names is limited to NCBI taxonomy, the combination of genomes into either group is not. As an example, users can combine *Escherichia* and *Shigella* genomes as one group and compare them against *Salmonella*. By default, each genome of group A will be compared with each genome of group B. However, an option is provided for users to also compare genomes within each group. Once the query is submitted, a table will be provided to show the genomic and marker gene identity metrics for all pairwise comparisons. Moreover, a 2D graph will be provided that allows users to view the relationship between any two metrics. Both the result table and the graph can be downloaded for users to conduct downstream analysis. Comparisons between all pairs of genomes are pre-conducted and available on our Web site.

By providing the comparison between and within user-defined genome groups, the POGO database can help answer various evolutionary biology questions. Figure 3 shows an example of POGO being used to compare genomes within the genus *Streptococcus*. The number of orthologs decreases as the average AAI decreases. Such comparison may be useful in studies of core versus pan genomes. Meanwhile similar analysis between different microbial genera may be helpful in studies of genome reduction, stability of gene content across lineages, etc.

AVERAGE RANKING OF MARKER GENES

As mentioned earlier, different marker genes may provide the best phylogenetic distinction within different lineages, and several recent studies have started looking into using markers other than the 16S rRNA gene for phylogenetic inference. Markers that are more slowly evolving (e.g. 16S rRNA) may be better for inferring more distantly related phylogenies. At the same time, faster evolving markers may provide better distinction when closely related microbes are considered. POGO allows users to compare different marker genes with regard to how fast they evolve within a group of interest. To this end, POGO can be used to rank markers according to how fast they evolve within a group of genomes (provided that the selected genomes contain all marker genes). Additionally, POGO provides the average ranking of each marker gene across all pairs of genomes queried by user (Figure 2). The average ranking reflects how fast-evolving a marker is compared with all other markers within a selected set of genomes. Markers that are ranked higher are faster evolving than those ranked lower.

One example of using the average ranking table is shown in Figure 4, which was generated from genome

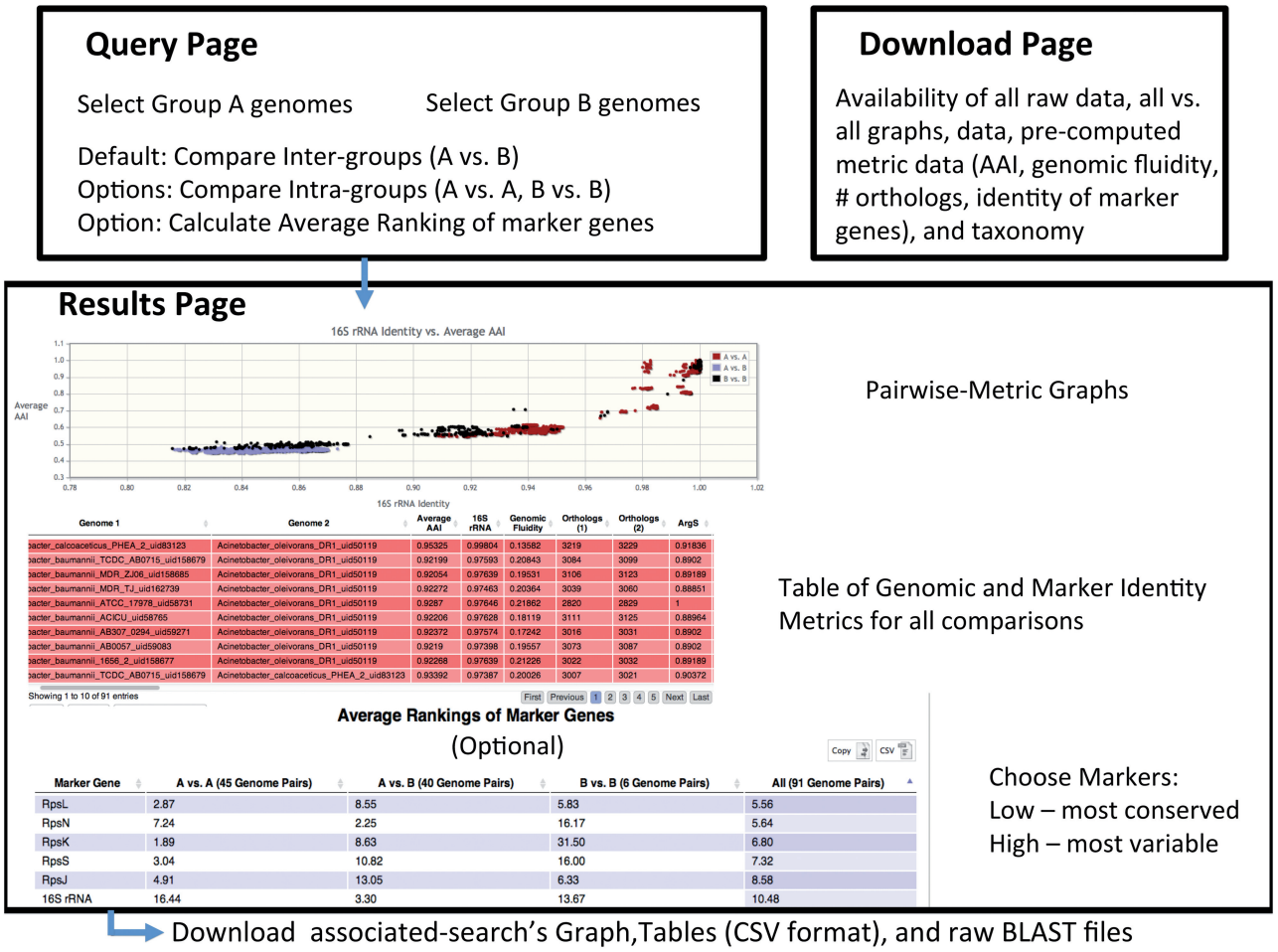

Figure 2. Content of the POGO database. The query page allows users to select genomes of interest into two lists. The results and download options based on the selections are shown on the results page. A separate download page exists for users who wish to download 'raw' data *en masse*.

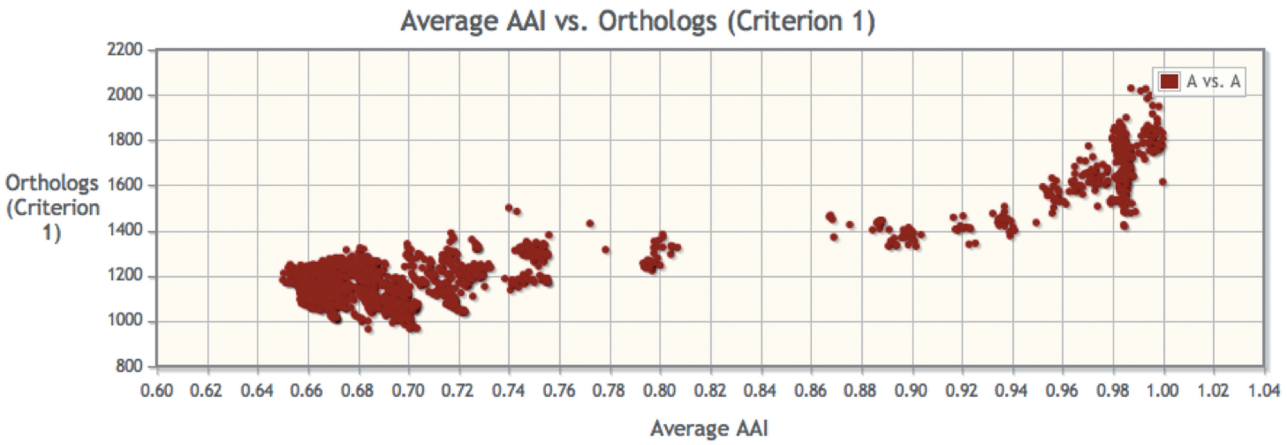

Figure 3. Example of a result graph of pairwise genome comparisons within the genus *Streptococcus*. POGO can be used to visualize that *Streptococcus* may best be separated into two groups where genomes that have >85% average AAI share >1400 orthologs in general, whereas genomes that have <85% average AAI share <1400 orthologs in general.

comparisons within the genus *Bacillus*. The gene *coaE* is the highest ranked (least conserved) among all marker genes, whereas the 16S rRNA is the lowest ranked gene (most conserved). For genomes with an average AAI >75%, we see that the 16S rRNA gene identity correlates poorly to the average AAI, whereas the *coaE* gene correlates well to the average AAI. We also generated three unweighted pair group method with arithmetic mean

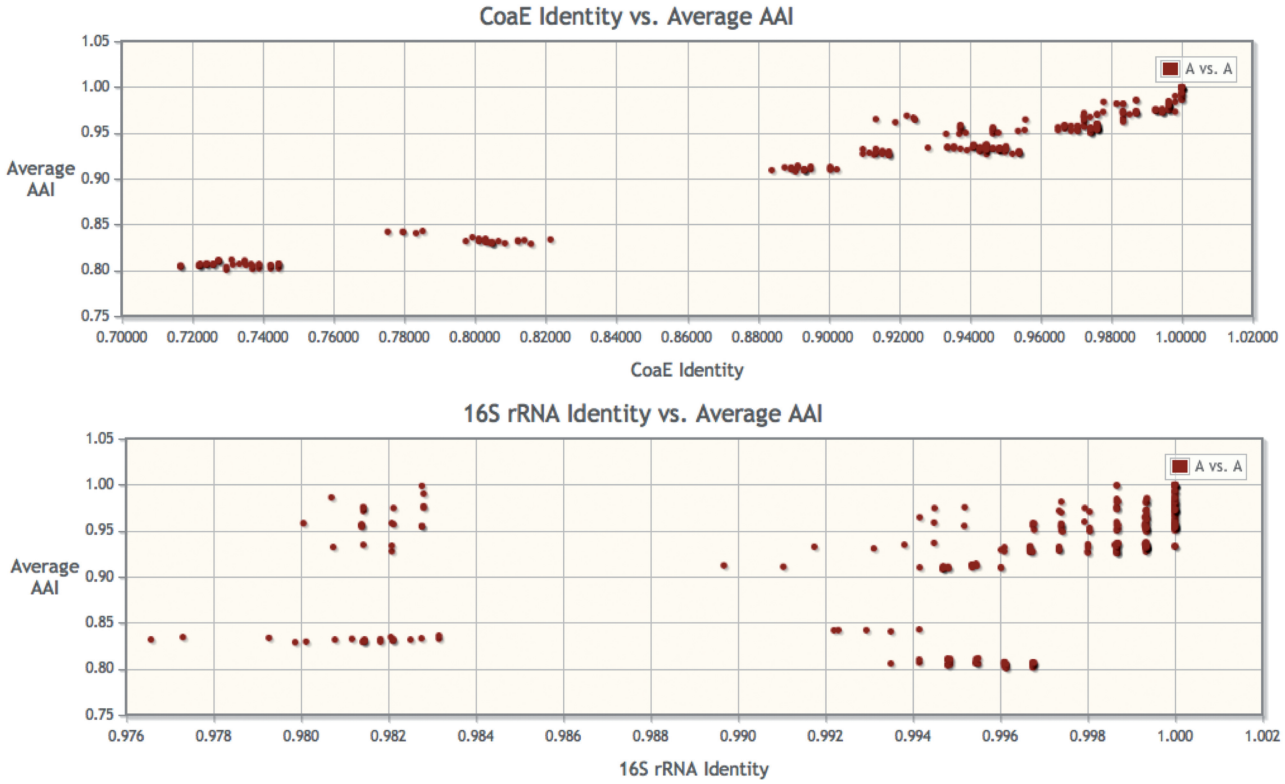

Figure 4. Using POGO-DB to correlate the percentage identity of marker genes to AAI. In this example, we used POGO-DB to correlate the percentage identities of both *coaE* and 16S rRNA with AAI for all *Bacillus* pairwise comparisons in which AAI was >75%. For these comparisons, the percentage identity of 16S rRNA is poorly correlated to the average AAI (Spearman correlation $rho = 0.61$, $P \ll 0.001$). In contrast, the percentage identity of the *coaE* gene, which is the fastest evolving marker gene within *Bacillus* according to POGO's average ranking feature, correlates much better with the average AAI (Spearman correlation $rho = 0.94$, $P \ll 0.001$).

(UPGMA) phylogenetic trees of *Bacillus* using data of AAI, percentage identity of 16S rRNA and percentage identity of *coaE* (Supplementary Figures S1–S3). We found that the tree generated using *coaE* much better resembles the AAI tree, compared with the tree generated by 16S rRNA gene [split distance computed using the phylogenetic tree comparison software TOPD/FMTS (30) is 0.340 and 0.638, respectively]. Therefore, *coaE* may be a better marker than the 16S rRNA for highly similar *Bacillus* genomes. Such information should be of great use to scientists embarking on metagenomic studies in which they are particularly interested in discriminating between members of closely related lineages and/or inferring a finer-grained phylogeny than allowed by analyses of the 16S rRNA gene.

DATABASE IMPLEMENTATION

Overall structure

The POGO Web site was written using PHP, JavaScript and HTML5. The use of these platforms allows for an interactive user experience with dynamic graphs, tables and charts. When the user queries our database, the Apache web stack communicates with our MySQL database to efficiently search for the requested genomic comparisons.

Query/home page

The home page allows users to select genomes and add them to either the group A or group B. Users can search for genomes to study using genus, species and strain names. It should be noted that all names are consistent with the NCBI taxonomy nomenclature; for entries not assigned a species or genus name in the NCBI taxonomy, we classified them into the category of 'Unidentified'. For example, genome *Clostridium_difficile_630_uid57679* can be found under tag 'Unidentified (genus)' –> '[Clostridium] difficile (species)'. However, a user can add any genome or genomes to group A or group B for both inter- and intra-group comparison, regardless of taxonomy boundaries. This benefits the users in scenarios in which the NCBI taxonomic classification does not serve their needs. For example, strains of *Shigella* are known to be, for all intents and purposes, members of the *E. coli* species (31), and users may want to group such strains together.

Once users place genomes into groups, they can choose whether to compare all genomes in group A with all those in group B (default), and can also choose to carry out comparisons of all genomes within A against each other, and/or of all organisms within B against each other. Users can do all three of these at once, in which case the results will be color-coded according to whether a certain dot on

the graph represents and A to A, A to B or B to B comparison. Users can also choose the option to rank marker genes. If users are interested in comparing two groups, they need to place genomes into both group A and group B, but if they just wish to perform intra-group comparisons, they can place genomes in only one of the two groups.

Results page

When users submit their query, our Web site gathers the requested comparisons and returns them in a result scatter plot, and also provides a table containing the same data. If the users selected to see the average marker ranking, it shows a table of the rankings below the result chart. On this page, users can select different metrics to display on the graph axis, download the graph and zoom on the graph to get a closer view. The user can also hover over points and see more detailed information about each.

In the Results page, users can easily copy the result chart or the average ranking chart to clipboard, download a comma-separated value file containing the metrics from the charts or download the scatter plot. The users can also select a subset or all of the comparisons they have conducted and download the corresponding raw BLAST files in a tar archive.

Download page

On the download page, users can download all the comparison data and BLAST files that POGO contains. These include the 16S rRNA pairwise identity for all genomes, the genome and marker gene identities for all pairs of genomes we analyzed, figures pre-drawn containing all pairwise comparisons, output files of all whole-genome BLAST comparisons we performed and the NCBI taxonomy annotation applied on POGO.

API access

A key feature of POGO is the ability for users to directly query our database. Using our API, we provide users with a way to access data more efficiently. We recognize that users may have different needs than what our Web site provides, so providing a different way to access the information is important. We created this API because users may want to feed the data through a pipeline, download large sets of data or manipulate them in other ways. A user accesses the API's URL with certain parameters and is returned a JavaScript Object Notation (JSON) or comma-separated value file. This also allows users to perform more complex queries than what our Web site interface allows. For example, users can gather all genome pairs where the 16S rRNA gene identity is >99% and the *valS* gene identity is <20% at the same time, or comparisons containing the genus *Bacillus*, where the number of orthologs is >800. More information about how to use the API can be found on our site (http://pogo.ece.drexel.edu/api_doc.php).

Database updates

We plan to update POGO-DB on a yearly basis, incorporating newly sequenced genomes as computational feasibility permits.

Database limitations

It should be noted that there are some limitations to the database. Although such limitations should be considered, they are minor in respect to the power of the database. As mentioned in the previous paragraph, some taxonomic selections must be made under the 'Unidentified' category, as we rely on NCBI taxonomy and some taxa still remain un-annotated or in-transition for the genus and species levels. Second, pairwise BLAST comparisons were only conducted for genomes that had at least 80% maximum (the highest value was taken if more than one 16S rRNA was available) 16S rRNA identity. The user may note that if not all combinations of pairwise genomes are available in the result table, it is due to the fact that the genomes are so distant that their best matching 16S rRNA genes are <80% identical. Finally, it is important to note that we relied on NCBI gene calling in all our analyses. Therefore, our estimates of gene fluidity and number of orthologs are reliant on these annotations.

DISCUSSION AND PERSPECTIVES

POGO allows users to not only compare genomes and marker genes but to also visualize and summarize this information in a graphical easy-to-use interface while also making the vast information downloadable for flexible analysis. Information provided by POGO-DB, with a few simple mouse clicks can help researchers navigate the ill-defined world of taxonomy and allow users to interactively explore microbial evolution. For example, if a user is interested in understanding whether intra-lineage variation is higher in lineage A compared with lineage B, it would take only minutes to find out using POGO-DB. It would also take only minutes to understand how different aspects of similarity (e.g. sequence similarity versus gene content similarity) relate to each other within a lineage or lineages of interest.

In addition, POGO-DB allows users to easily find marker genes that, within their lineages of interest, best correlate with overall average AAI. Overall, AAI should be the best sequence-based metric of phylogenetic relatedness, as it encompasses information in a genome-wide manner from a large number of genes, rather than rely on only a single or few loci. However, it is impossible to obtain data of AAI from metagenomic studies where only small fragments of DNA from each bacterium can be sequenced. By allowing users to find the markers that best correlate with overall AAI, we allow them to identify the marker that best predicts phylogenetic relatedness within lineages of interest. Users can also use POGO-DB data to construct phylogenetic trees based on pairwise similarity of different markers and compare them with trees built based on AAI. This should allow users to select the markers that best reconstruct the phylogeny of

the lineage in which they are interested. Therefore, POGO-DB provides users with the needed data to select the best markers to use in their metagenomic studies. Finally, POGO-DB allows users to easily rank marker genes within a lineage or lineages to ascertain which markers are evolving faster and which are slower within those lineages. Markers that are more slowly evolving (e.g. the 16S rRNA) may be better for inferring more distantly related phylogenies. After all, variation within such slowly evolving markers is less likely to be saturated. At the same time, faster evolving markers may provide better distinction when closely related microbes are considered.

As an example, we have used data of marker ranking from POGO-DB to show that within *Bacillus* the 16S rRNA gene is the slowest evolving marker, whereas the *coaE* gene is the fastest. In Figure 4, we present data from POGO-DB showing that the percentage identity of the *coaE* gene correlates much better with AAI than the percentage identity of the 16S rRNA gene. We also generated three phylogenetic trees of *Bacillus* and found that the *coaE* tree resembles the AAI tree much better than the 16S rRNA tree (Supplementary Figures S1–S3). Therefore, we can conclude, based solely on data that are easily accessible on POGO-DB, that *coaE* infers the phylogeny of the *Bacillus* lineage much more accurately than 16S rRNA. This exemplifies how useful POGO-DB can be for selecting marker genes for specific lineages of interest.

The ability to choose appropriate lineage-specific marker genes will have a wide impact on fields of microbial source tracking (32,33), investigating molecular mechanisms and its impact on microbial evolution (34) and accurate phylogenetic reconstruction of the microbial diversity in ecosystems (35).

In conclusion, POGO-DB provides users an easy manner in which to compare several aspects of genome relatedness, and to identify marker genes that best recapitulate genome relatedness. In addition, we provide users with raw BLAST results that took great computational resources to generate. This will allow users full flexibility to conduct any downstream analyses they conceive. POGO-DB will therefore be of great use to anyone interested in studying prokaryote genome evolution, and/or in choosing the best phylogenetic marker genes for their metagenomic studies.

AVAILABILITY

The web interface is available at http://pogo.ece.drexel.edu. The data (in JSON format) can be queried with PHP via: http://pogo.ece.drexel.edu/query.php (see the API documentation for more information at http://pogo.ece.drexel.edu/api_doc.php). Raw data can be downloaded via http://pogo.ece.drexel.edu/downloads.php.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation [OCI-1053575]. The authors are grateful to the XSEDE startup resources for this opportunity. Research by R.H. is carried out in the Rachel & Menachem Mendelovitch Evolutionary Process of Mutation & Natural Selection Research Laboratory.

FUNDING

Funding for open access charge: This work was initiated by the Louis and Bessie Stein Family Fellowship, supporting exchanges between Drexel and Israeli Universities. The work performed by the Drexel Ecological and Evolutionary Signal-processing and Informatics (EESI) laboratory was supported in part by a National Science Foundation (NSF) CAREER award number [0845827], NSF award number [1120622] and Department of Energy (DOE) Office of Science (BER) award [DE-SC0004335]; R.H. is Supported by an ERC FP7 CIG grant [N° 321780], by a Yigal Allon Fellowship awarded by the Israeli Council for Higher Education and by the Robert J. Shillman Career Advancement Chair.

Conflict of interest statement. None declared.

REFERENCES

1. Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol. CB*, **11**, 941–950.
2. Raymond,J., Zhaxybayeva,O., Gogarten,J.P., Gerdes,S.Y. and Blankenship,R.E. (2002) Whole-genome analysis of photosynthetic prokaryotes. *Science*, **298**, 1616–1620.
3. Uchiyama,I., Mihara,M., Nishide,H. and Chiba,H. (2013) MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. *Nucleic Acids Res.*, **41**, D631–D635.
4. Bohlin,J., Snipen,L., Cloeckaert,A., Lagesen,K., Ussery,D., Kristoffersen,A.B. and Godfroid,J. (2010) Genomic comparisons of Brucella spp. and closely related bacteria using base compositional and proteome based methods. *BMC Evol. Biol.*, **10**, 249.
5. Riesenfeld,C.S., Schloss,P.D. and Handelsman,J. (2004) Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.*, **38**, 525–552.
6. Gilbert,J.A., Meyer,F., Jansson,J., Gordon,J., Pace,N., Tiedje,J., Ley,R., Fierer,N., Field,D., Kyrpides,N. *et al.* (2010) The earth microbiome project: meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory 6 October 2010. *Stand. Genomic Sci.*, **3**, 249–253.
7. Gevers,D., Knight,R., Petrosino,J.F., Huang,K., McGuire,A.L., Birren,B.W., Nelson,K.E., White,O., Methe,B.A. and Huttenhower,C. (2012) The human microbiome project: a community resource for the healthy human microbiome. *PLoS Biol.*, **10**, e1001377.
8. Shade,A., Gregory-Caporaso,J., Handelsman,J., Knight,R. and Fierer,N. (2013) A meta-analysis of changes in bacterial and archaeal communities with time. *ISME J.*, **7**, 1493–1506.
9. Kittelmann,S., Seedorf,H., Walters,W.A., Clemente,J.C., Knight,R., Gordon,J.I. and Janssen,P.H. (2013) Simultaneous amplicon sequencing to explore co-occurrence patterns of bacterial, archaeal and eukaryotic microorganisms in rumen microbial communities. *PLoS One*, **8**, e47879.
10. Pei,A.Y., Oberdorf,W.E., Nossa,C.W., Agarwal,A., Chokshi,P., Gerz,E.A., Jin,Z.D., Lee,P., Yang,L.Y., Poles,M. *et al.* (2010)

Diversity of 16S rRNA genes within individual prokaryotic genomes. *Appl. Environ. Microb.*, **76**, 3886–3897.
11. McDonald,D., Price,M.N., Goodrich,J., Nawrocki,E.P., DeSantis,T.Z., Probst,A., Andersen,G.L., Knight,R. and Hugenholtz,P. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.*, **6**, 610–618.
12. Cole,J.R., Wang,Q., Cardenas,E., Fish,J., Chai,B., Farris,R.J., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Marsh,T., Garrity,G.M. *et al.* (2009) The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.
13. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glockner,F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
14. Vos,M., Quince,C., Pijl,A.S., de Hollander,M. and Kowalchuk,G.A. (2012) A Comparison of rpoB and 16S rRNA as markers in pyrosequencing studies of bacterial diversity. *PLoS One*, **7**, e30600.
15. Vetrovsky,T. and Baldrian,P. (2013) The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One*, **8**, e57923.
16. Achenbach,L.A., Carey,J. and Madigan,M.T. (2001) Photosynthetic and phylogenetic primers for detection of anoxygenic phototrophs in natural environments. *Appl. Environ. Microb.*, **67**, 2922–2926.
17. Walsh,D.A., Bapteste,E., Kamekura,M. and Doolittle,W.F. (2004) Evolution of the RNA polymerase B ' subunit gene (rpoB ') in Halobacteriales: a complementary molecular marker to the SSU rRNA gene. *Mol. Biol. Evol.*, **21**, 2340–2351.
18. Case,R.J., Boucher,Y., Dahllof,I., Holmstrom,C., Doolittle,W.F. and Kjelleberg,S. (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl. Environ. Microb.*, **73**, 278–288.
19. Segata,N. and Huttenhower,C. (2011) Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. *PLoS One*, **6**, e24704.
20. Schloissnig,S., Arumugam,M., Sunagawa,S., Mitreva,M., Tap,J., Zhu,A., Waller,A., Mende,D.R., Kultima,J.R., Martin,J. *et al.* (2013) Genomic variation landscape of the human gut microbiome. *Nature*, **493**, 45–50.
21. Wu,M. and Eisen,J.A. (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.*, **9**, R151.
22. Liu,B., Gibbons,T., Ghodsi,M., Treangen,T. and Pop,M. (2011) Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics*, **12(Suppl. 2)**, S4.
23. Konstantinidis,K.T. and Tiedje,J.M. (2005) Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.*, **187**, 6258–6264.
24. Kislyuk,A.O., Haegeman,B., Bergman,N.H. and Weitz,J.S. (2011) Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics*, **12**, 32.
25. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
26. Mount,D.W. (2007) Using the basic local alignment search tool (BLAST). *CSH Protoc.*, **2007**, pdb top17.
27. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.
28. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
29. Kristensen,D.M., Kannan,L., Coleman,M.K., Wolf,Y.I., Sorokin,A., Koonin,E.V. and Mushegian,A. (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, **26**, 1481–1487.
30. Puigbo,P., Garcia-Vallve,S. and McInerney,J.O. (2007) TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics*, **23**, 1556–1558.
31. Hershberg,R., Tang,H. and Petrov,D.A. (2007) Reduced selection leads to accelerated gene loss in *Shigella*. *Genome Biol.*, **8**, R164.
32. Shanks,O.C., Domingo,J.W., Lu,J., Kelty,C.A. and Graham,J.E. (2007) Identification of bacterial DNA markers for the detection of human fecal pollution in water. *Appl. Environ. Microbiol.*, **73**, 2416–2422.
33. Medalie,L., Matthews,L.J. and Stelzer,E.A. (2011) Using host-associated genetic markers to investigate sources of fecal contamination in two Vermont streams. *USGS Sci. Invest. Rep.*, **30**, 2011–5113.
34. Arber,W. (2000) Genetic variation: molecular mechanisms and impact on microbial evolution. *FEMS Microbiol. Rev.*, **24**, 1–7.
35. Torsvik,V. and Ovreas,L. (2002) Microbial diversity and function in soil: from genes to ecosystems. *Curr. Opin. Microbiol.*, **5**, 240–245.